Modelli statistici per l'interpretazione dei dati LANDSAT(*)

Agostino Tarsitano
Università degli studi della Calabria
Dipartimento di economia e Statistica
87030 Arcavacata di Rende (Cs)
agotar@unical.it

Riassunto.
In questo lavoro si discute la classificazione non supervisionata applicata alla conoscenza del territorio. In particolare viene presentata la tecnica ISODATA per clusters a metrica variabile. La tecnica è stata impiegata per la classificazione di dati rilevati a distanza con scanner multispettrali in Calabria trovandola abbastanza pratica sebbene non ancora pronta per essere utilizzata regolarmente come strumento di routine per la quantificazione delle risorse agricole.

Keywords: Isodata, Cluster analysis, valori remoti, componenti principali, dati multispettrali

(*) Lavoro presentato al convegno "Distribuzione e quantificazione delle risorse territoriali in Calabria tramite telerilevamento e sistemi di controllo a terra". Centro Calabrese Documentazione e Studi. Cosenza, 27 ottobre 1979

1. Introduzione

Nel maggio del 1979 si è avviato presso il Dipartimento di Ecologia dell'Università della Calabria un progetto di ricerca avente come oggetto il telerilevamento da satellite come strumento di inventario, controllo e gestione delle risorse territoriali.

Gli obiettivi del progetto nel lungo periodo erano da una parte lo sviluppo di una metodologia nuova per l'inventario delle risorse naturali per le condizioni calabresi; d'altra parte studiare la fattibilità tecnica ed economica di questi metodi. Il progetto è infatti volto sia alla applicazione che allo studio e revisione delle tecniche di base sviluppate nei centri sul *remote sensing* statunitensi. Nel breve periodo gli obiettivi erano più limitati e più che altro orientati ad un contatto preliminare con le tecniche già stabilizzate in questo campo. Fra questi traguardi più immediati era prevista un'indagine conoscitiva sugli strumenti quantitativi condensata in un *package* di programmi predisposto per la trattazione dei dati MSS (MultiSpectral Scanner). La relazione presente riguarda in maniera precipua gli aspetti statistico-matematici del telerilevamento.

In sostanza, questi aspetti quantitativi possono, approssimativamente, essere inquadrati in due tipi di approccio: quello *supervised* e quello *unsupervised*. Nel primo approccio vengono prima selezionate, per una zona test, delle aree di ampiezza significativa e la cui copertura è del tutto nota (*training areas*). Utilizzando le informazioni qui ricavate vengono stimati parametri delle distribuzioni normali multivariate che si suppone siano associate ad ogni particolare tipo di copertura. Successivamente si passa alla classificazione dei dati relativi all'intera zona test. Nell'altro approccio, quello *unsupervised*, vengono utilizzate delle tecniche di *cluster analysis* che dividono i dati della zona test in un certo numero di gruppi spettrali distinti. A questo punto è necessario identificare quale tipo di copertura è rappresentato da ciascun gruppo spettrale.

Le differenze fra i due approcci sono più apparenti che reali. Sebbene nel *supervised* si operi prima nelle aree di *training* e dopo nella zona test, mentre nell'*unsupervised* si affronta subito la zona test, di fatto, considerando quest'ultima come area di *training* o viceversa la differenza "dimensionale" scompare. In seconda istanza i modelli statistico-matematici che vengono usati sono anche abbastanza simili, si diversificano però a causa della discriminante, unica se si vuole, che passa fra i due approcci: il tempo di utilizzo delle informazioni disponibili sulla copertura delle zone in esame. Contemporaneamente alle elaborazioni nel primo approccio, ad elaborazioni ultimate nel secondo. In entrambi comunque il problema cruciale è la generalizzabilità dei risultati ottenuti nelle aree di *training* alla zona test e da questa poi ad una

dimensione territoriale più ampia e ad una loro proiezione nel tempo. In questo senso è possibile individuare due caratteristiche fondamentali che è necessario riscontrare in ogni progetto di utilizzo del remote sensing per fini di pianificazione e programmazione territoriale: Interdisciplinarietà ed Iteratività.

Interdisciplinarietà in quanto gli elementi di dubbio coinvolti in questo tipo di ricerche provengono dalle fonti più diverse ed in un certo senso fra di loro lontane. E' indispensabile quindi che gli studiosi afferenti alle varie discipline interessate si sforzino di trovare un'area comune, un'intersezione di metodi che congiuntamente consentano di addivenire ad un uso maggiormente esteso e semplificato di uno strumento di conoscenza del territorio validissimo, quale il telerilevamento. Iteratività in quanto non è pensabile che la pedissequa esecuzione di fasi (così si presentano oggi gli studi sul *remote sensing*) possa portare a risultati apprezzabili. Sembra invece più plausibile una ripetitività sequenziale dei momenti dell'analisi: rilevamento a terra, studio delle immagini (da ogni piattaforma), elaborazione dei dati (secondo entrambi gli approcci), verifica dei risultati ottenuti. La iteratività va comunque intesa in modo completo ma non rigido e che pure lasci spazio a pause di riflessione e rimeditazione. Le ricerche in cui più si è spinto sulla interazione e sulla ciclicità sono quelle che più hanno contribuito, per la qualità degli obiettivi raggiunti, all'attuale successo del remote sensing, successo attualmente accolto solo nelle Università e nei centri di ricerca specializzati.

Nel lavoro presente vengono esposte le esperienze avute nel seguire l'approccio *unsupervised* particolarmente privilegiato nell'impostazione della ricerca che si è condotta. Nel secondo paragrafo viene discusso il modello teorico della classificazione *unsupervised* mentre nel terzo si tratta dell'algoritmo a cui esso dà luogo. Infine, nel quarto paragrafo, sono studiati due importanti problemi del *remote sensing* comuni ad entrambi gli approcci: Riduzione della dimensionalità e valori remoti.

2. Un modello per la classificazione *unsupervised.*

La parte essenziale di un progetto realistico per l'analisi di dati da *remote sensing* è l'applicazione delle tecniche di *cluster analysis* che riescono a dare una estesa base informativa per la classificazione dei dati. Compito di queste tecniche è di decomporre i dati di partenza in un certo numero di gruppi (o *cluster*) tali che le unità appartenenti ad uno stesso gruppo siano "più simili" - conformemente ad una data misura di similarità- che non unità appartenenti a gruppi differenti. Il numero di gruppi e le loro caratteristiche sono da deter-

minarsi. La costruzione dei gruppi è di solito condotta iterativamente ed in modo da soddisfare ad un qualche criterio di ottimizzazione. I gruppi risultanti dovrebbero infine corrispondere agli aspetti più marcati presenti nei dati. Sulla scia dello sviluppo dei computer e delle tecniche di programmazione si è assistito negli ultimi tempi ad un fiorire di tecniche di *clustering* tanto che oramai per ogni particolare applicazione è disponibile una tecnica di *clustering* altrettanto particolare. Anche per la classificazione *unsupervised* di dati Landsat esistono procedure sviluppate *ad hoc* (Ball e Hall, 1967) che tuttavia risentono della mancanza di un assetto teorico ben definito.

Nella predisposizione di un modello per la classificazione si è in questo senso adottata la formulazione probabilistica esposta in Scott e Symons, 1971 che vede la *cluster analysis* come: *"... A form of mixtures analysis for finite mixtures of multivariate distributions"*. Questo approccio offre il vantaggio di fornire una premessa di tipo più formale ad una analisi, quale quella di *cluster*, che altrimenti rischia di rimanere nulla più di un bel gioco con i numeri. Nel selezionare però le ipotesi che precisano il modello probabilistico si dovrà necessariamente tenere conto della realtà particolare che esso deve rappresentare. E' opportuno quindi per prima cosa mettere su una esposizione sintetica, ma completa del ruolo che ha, nel telerilevamento, la classificazione *unsupervised*, che possa servire come base di riferimento alle successive astrazioni del modello.

Si ha di fronte una "scena" che consiste in una regione rettangolare con *r* righe (*scanlines*) e *c* colonne di elementi di risoluzione (*pixel*). Ad ogni elemento è associato un vettore *X (px1)* di misurazioni relative alla digitalizzazione delle risposte spettrali nei *p* canali ottici disponibili sul satellite. Sulla scena esistono esattamente *G* categorie C_g, $g=1,2,\ldots,G$ dove per categoria di intende un ente al terreno (coltivazione, acque, urbanizzazioni, etc.) presente in misura tale da riuscire ad impressionare un pixel (59x69)m. Ogni categoria è individuata da un particolare insieme di risposte spettrali (una per ogni canale). Chiaramente la combinazione di risposte associata ai pixels di una data categoria non rimane costante (ad eccezione forse di enti come ad esempio il mare) su tutta la scena. L'energia trasmessa dalla superficie terrestre ad un *remote sensor* è soggetta a dispersione a causa delle particelle sospese nell'atmosfera, dell'assorbimento del vapore acqueo, del diossido di carbonio e dell'ozono. In effetti le "*spectral signature*" registrate dai sensori sono sempre attenuate e modificate dagli effetti dell'atmosfera. Sono anche possibili sfasamenti dell'energia trasmessa attraverso l'atmosfera; sia in senso temporale che spaziale.

Le correzioni radiometriche e geometriche che vengono applicate ai dati grezzi non possono fare altro che ridurre gli effetti di disturbo rimane tutta-

via una certa dose di “casualità” nella registrazione dei livelli di radianza. La variabilità delle risposte dei *pixel* di una categoria è però soprattutto determinata dalle caratteristiche della scena a cui i *pixels* stessi si riferiscono. In realtà esistono molte variabili che influenzano la risposta spettrale degli enti al terreno (particolarmente per le categorie vegetative) quali ad esempio: l’illuminazione, la morfologia, il punto di vista colto dai sensori ed altre ancora.

Possiamo tenere conto globalmente degli effetti di disturbo ipotizzando per ogni categoria al terreno una distribuzione probabilistica normale p-variata:

$$\phi\left(\boldsymbol{X},\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g\right)=(2\pi)^{-m/2}\left|\boldsymbol{\Sigma}_g\right|^{-1/2}\exp\left\{-0.5\left(\boldsymbol{X}-\boldsymbol{\mu}_g\right)^t\Sigma_g^{-1}\left(\boldsymbol{X}-\boldsymbol{\mu}_g\right)\right\};\ g=1,2,\ldots,G \qquad (2.1)$$

dove μ_g è il vettore delle medie delle risposte nei *p* canali e $\sum_g$ è la matrice di varianze-covarianze. Il vettore μ_g va inteso come combinazione specifica della categoria C_g ed a questo proposito si ipotizza che

$$\mu_i = \mu_j \text{ se e solo se } i=j \qquad (2.2)$$

Tale ipotesi è l’espressione di una completa distinguibilità spettrale fra le categorie in cui si è suddiviso il terreno nel senso che, se fosse possibile osservare l’ente g-esimo senza alcun disturbo, ogni *pixel* presenterebbe la combinazione di risposte μ_g e questa sarebbe diversa nelle diverse categorie presenti.

Di fianco all’ipotesi di distinguibilità spettrale delle categorie viene collocata quella di omogeneità che riguarda più direttamente la matrice $\sum_g$. Per ogni categoria C_g si ipotizza che

$$|\sigma_{gij}| < \infty \qquad (2.3)$$

dove σ_{gij} è l’elemento generico della matrice $\sum_g$. La (2.3) implica che $\sum_g$ è costituita da elementi finiti, ma va anche pensata come un limite, specificabile solo empiricamente, posto agli scostamenti dei pixel dai valori tipici in cui la categoria a cui appartengono in modo da dare alle categorie stesse una loro precisa connotazione spettrale. Si ipotizza inoltre che ognuna delle *G* categorie sia presente in N_g elementi di risoluzione, con N_g tale che

$$N_g > p,\ g=1,2,\ldots,G \qquad (2.4)$$

Questa condizione riguarda immediatamente la stimabilità dei parametri μ_g e $\sum_g$ ma va anche considerata in un senso più ampio. L'area occupata dalla categoria g-esima deve essere una percentuale apprezzabile dell'area totale della scena. In altre parole, vengono escluse dalla procedura di *clustering* le classi minori che, oltre a complicare la valutazione dei risultati, deviano i termini stessi dell'analisi. La 2.4 è affiancata dall'ipotesi di non singolarità delle matrici di varianze e covarianze

$$|\sum_g| \neq 0, \qquad g=1,2,\ldots,G \tag{2.5}$$

Sulla 2.5 il discorso verrà approfondito nel paragrafo 4.

Gli *(rxc)* elementi di risoluzione della scena costituiscono un campione casuale $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$ di osservazioni p-dimensionali. Ogni $\boldsymbol{X}_i$ può provenire da una ed una sola delle *G* popolazioni (categorie) $N_p(\boldsymbol{\mu}_g \boldsymbol{\Sigma}_g);\ g = 1,2,\ldots,G$. Assumiamo inoltre l'esistenza di un vettore di parametri di classificazione $\gamma=\{\gamma_1, \gamma_2,\ldots,\gamma_N\}$ tale che, per ogni dato $\boldsymbol{X}_i$ si abbia

$$\gamma_i = g \text{ se e solo } \boldsymbol{X}_i \in \mathrm{C_g} \tag{2.6}$$

Possiamo a questo punto scrivere la funzione di verosimiglianza per tutto il *set* di parametri coinvolti nel modello

$$L\left(\boldsymbol{\gamma}; \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_g; \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_g\right) = \prod_{g=1}^{k} \prod_{\gamma_r = g} \phi\left(\boldsymbol{X}_r, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right) \tag{2.7}$$

dove $\boldsymbol{X}_\gamma$ sta ad indicare che la osservazione $\boldsymbol{X}$ appartiene alla categoria individuata dal valore di γ_r (elemento generico di γ). Per un fissato valore di γ possiamo ottenere le stime di massima verosimiglianza per $\mu_1, \mu_2, \ldots, \mu_G$ differenziando il logaritmo della (2.7). In particolare, per la categoria g-esima si ha

$$\hat{\boldsymbol{\mu}}_g = \frac{1}{n_g} \sum_{\gamma_r = g} \boldsymbol{X}_r; \qquad \hat{\boldsymbol{\Sigma}}_g = \frac{1}{n_g - 1} \sum_{\gamma_r = g} \left(\boldsymbol{X}_r - \hat{\boldsymbol{\mu}}_g\right)\left(\boldsymbol{X}_r - \hat{\boldsymbol{\mu}}_g\right)^t \tag{2.8}$$

Sia per μ_g che $\sum_g$ sono naturalmente delle funzioni di γ e questo va precisato dato che non è esplicito nella notazione. Sostituendo le (2.8) nella (2.7) la funzione di verosimiglianza diventa una funzione solo di γ.

$$L(\boldsymbol{\gamma}) = -0.5\sum_{g=1}^{G}\sum_{\gamma_r=g}\left[\left(\boldsymbol{X}_r - \hat{\boldsymbol{\mu}}_g\right)^t \hat{\Sigma}_g^{-1}\left(\boldsymbol{X}_r - \hat{\boldsymbol{\mu}}_g\right) + Ln\left(\hat{\boldsymbol{\Sigma}}_g\right) + 2Ln(G)\right] \qquad (2.9)$$

Consegue che la stima di massima verosimiglianza $\hat{\boldsymbol{\gamma}}$ per γ può essere ottenuta scegliendo la partizione delle *N* osservazioni in *G* gruppi che massimizza la (2.9).

Il modo più sicuro di determinare il massimo della (2.9) sarebbe quello di valutarla su tutte le possibili partizioni di *N* elementi in *G* gruppi che è però fuori di ogni praticabilità quando si elabora una mole di dati anche solo moderatamente cospicua quale quella di una scena Landsat. Si deve allora necessariamente ricorrere ad un algoritmo di ottimizzazione che iterativamente tenti di pervenire alla massimizzazione della (2.9). A questo argomento è dedicato il paragrafo che segue.

3. Un algoritmo per la classificazione *unsupervised*

La procedura che si è adottata è simile, nei principi, alla nota tecnica I.S.O.D.A.T.A. (*Iterative Self-Organizing Data Analysis Technique A*) proposta in Ball e Hall (1967) e di frequente richiamata nell'analisi di dati Landsat. Si è comunque preferito collocare la tecnica nell'ambito del modello probabilistico tracciato nel paragrafo 2 in modo da avere un riferimento meno soggettivo e più sicuro.

Una volta fissati due interi *H* e *G* dove *H* indica il numero di *clusters* iniziali da formare e *G* il numero di *clusters* che si ipotizza siano presenti sulla scena in esame, la tecnica sviluppata ha tre momenti distinti:

a) La formazione dei *cluster*s iniziali.

b) L'assegnazione delle unità ai *clusters* già informati.

c) La revisione del numero corrente di *clusters* volta ad arrivare da *H* a *G*.

Partizione iniziale

Si selezionano a caso *H* unità dell'insieme dei dati e vengono poste come centroidi μ_h, $h= 1,\ldots, H$ degli *H clusters* iniziali. L'assegnazione delle unità è fatta, in questa fase iniziale, in base alla regola.

$$\gamma_r = m \Rightarrow \left(\boldsymbol{X} - \boldsymbol{\mu}_m\right)^t \Sigma^{-1}\left(\boldsymbol{X} - \boldsymbol{\mu}_m\right) = \underset{1\le h\le H}{Min}\left\{\left(\boldsymbol{X} - \boldsymbol{\mu}_h\right)^t \Sigma^{-1}\left(\boldsymbol{X} - \boldsymbol{\mu}_h\right)\right\} \qquad (3.1)$$

La 3.1 è legata al modello sviluppato nel paragrafo 2. nell'ipotesi che tutti i gruppi abbiano la stessa matrice di dispersione $\sum$. La (3.1) è una approssima-

zione piuttosto forte, ma consente, anche nella fase iniziale, di rimanere vicini al quadro teorico che si è adottato rispetto ad altri metodi che di fatto assumono come matrice di varianze e covarianze, unica per tutti i gruppi, la matrice Identità.

Assegnazione delle unità
L'assegnazione di una unità ad un particolare *cluster* dev'essere governata dall'obiettivo di massimizzazione la (2.9) o più formalmente dalla stima del vettore di classificazione γ. Questo implica che l'unità i-esima è assegnata al cluster g-esimo se e solo se

$$\left(\boldsymbol{X_i}-\boldsymbol{\mu}_g\right)^t \Sigma_g^{-1}\left(\boldsymbol{X_i}-\boldsymbol{\mu}_g\right)+n_g Ln\left(\left|\Sigma_g\right|\right)=\underset{1\le h\le H}{Min}\left\{\left(\boldsymbol{X_i}-\boldsymbol{\mu}_h\right)^t \Sigma_h^{-1}\left(\boldsymbol{X_i}-\boldsymbol{\mu}_h\right)+n_h Ln\left(\left|\Sigma_h\right|\right)\right\} \quad 3.2$$

In questo modo ognuna della N unità darà il contributo più basso alla (2.9) e dunque la (3.2) avvicina il valore corrente di γ alla stima di massima verosimiglianza di γ. Una volta che sia stata eseguita la (3.2) per tutte le N unità deve essere ripetuto il calcolo delle stime dei μ_g e $\sum_g$ dopo però aver eliminato quei *clusters* che comprendono un numero limitato di unità (il limite è fissato a priori ed in base a valutazioni più che altro soggettive si veda l'ipotesi 2.2).

Revisione del numero dei clusters
Quando l'algoritmo ha superato la fase di assegnazione e di rilocazione delle unità passa alla revisione del numero di *clusters*. Se il valore corrente di H è uguale a G le elaborazioni riprendono al punto secondo a meno che la partizione non abbia soddisfatto al "*clustering criterion*" scelto per orientare la stima di γ. Nel nostro caso si punta alla minimizzazione di $|\boldsymbol{W}|$ dove

$$\boldsymbol{W}=\sum_{\boldsymbol{h=1}}^{\boldsymbol{H}}\hat{\Sigma}_h \qquad (3.3)$$

Se invece $H<G$ allora un *cluster* sarà suddiviso in due. Occorre ovviamente scegliere quale *cluster* bipartire. Il candidato ideale in base agli obiettivi finali dell'algoritmo è quello per cui

$$\left|\hat{\Sigma}_m\right|=\underset{1\le h\le H}{Max}\left\{\left|\boldsymbol{\Sigma}_h\right|\right\} \qquad (3.4)$$

cioè si opera la suddivisione del *cluster* che presenta la più alta varianza ge-

neralizzata. E' anche importante determinare l'asse lungo il quale operare il sezionamento. Anche in questo caso la guida è la variabilità relativa. Sia p' la variabile tale che

$$\sqrt{\frac{\sum_{\gamma_r=m}\left(X_{rp'}-\hat{\mu}_{mp'}\right)^2}{n_m-1}}\Big/\hat{\mu}_{mp'}=\underset{1\le j\le p}{Max}\left\{\sqrt{\frac{\sum_{\gamma_r=m}\left(X_{rp'}-\hat{\mu}_{mj}\right)^2}{n_m-1}}\Big/\hat{\mu}_{mj}\right\} \tag{3.5}$$

i due nuovi *clusters* saranno formati collocando, diciamo nel primo, quelle unità per cui il valore nel canale p' è inferiore alla media. Se la differenza fra i centroidi dei nuovi *clusters* è piccola lo "*splitting*" non sarà effettuato.

Se infine $H>G$ allora due *clusters* saranno aggregati in uno. I *clusters* da aggregare sono quelli che presentano il valore più basso del test di Hotelling:

$$\left\{\left(\hat{\boldsymbol{\mu}}_i-\hat{\boldsymbol{\mu}}_j\right)^t\left[\left(\frac{n_i}{n_i+n_j-2}\right)\hat{\Sigma}_i+\left(\frac{n_j}{n_i+n_j-2}\right)\hat{\Sigma}_j\right]^{-1}\left(\hat{\boldsymbol{\mu}}_i-\hat{\boldsymbol{\mu}}_j\right)\right\}*\frac{\left(n_i+n_j-p-1\right)n_in_j}{p\left(n_i+n_j\right)\left(n_i+n_j-2\right)} \tag{3.6}$$

Il *lumping* dei due *clusters* non sarà effettuato se la statistica (3.6) risulta moderatamente elevata. Se H<>G e nè lo *splitting* e nè il *lumping* hanno avuto luogo le iterazioni vengono fermate.

Non ci sono naturalmente garanzie che la partizione finale ottenuta corrisponda alla stima di massima verosimiglianza di γ nel senso che la (2.9) può avere più massimi locali compresi con il massimo assoluto. Non ci sono inoltre garanzie che i *clusters* che fuoriescono come prodotto finale del processo prima descritto siano di interpretazione semplice e corrispondano alle idee intuitive che si avevano sul fenomeno a cui i dati fanno riferimento. Quello che si può fare è di provare l'algoritmo per diverse partizioni iniziali e per vari valori di *H* e di *G*. E' forse un modo di procedere lungo e costoso (specie se si tiene conto dei tempi di calcolo che l'algoritmo richiede per la sua esecuzione), ma è l'unico che può portare ad una disaggregazione plausibile della massa di dati di partenza.

Nella tabella 1 sono dati i risultati della *clustering* della scena selezionata come prima zona di *training* comprendente 8000 pixel. Il numero *H* di *clusters* iniziali è stato posto pari a 20 (tante erano le categorie che sembravano essere presenti sul terreno), in 5 prove con diversi valori di *G= 13,…,17* e per ogni dato valore di G, 5 prove per diversa aggregazione iniziale, quindi

Cluster	Pixels	$\lvert\Sigma_g\rvert$	Canale 4 μ e σ	Canale 5 μ e σ	Canale 6 μ e σ	Canale 7 μ e σ
1	564	0.1181+7	28.29	39.97	79.63	7.13
			3.49	5.94	4.63	5.97
2	365	0.1335+8	54.64	102.31	110.81	89.05
			7.39	6.93	6.75	7.73
3	731	0.7741+7	55.49	70.54	86.98	74.46
			5.83	9.69	8.22	6
4	773	0.2925+8	62.61	55.09	85.28	79.07
			7.49	9.52	12.73	10.97
5	426	0.4217+7	51.38	82.47	98.53	85.34
			6.89	5.52	5.12	5.21
6	248	0.1335+8	30.43	79.49	109.99	98.29
			6.82	5.99	5.85	5.7
7	506	0.2466+7	27.7	48.16	92.01	87.23
			2.7	6.36	5.2	6.27
8	786	0.4740+7	28.46	67.52	100.02	88.13
			3.54	8.56	11.1	12.8
9	398	0.3138+7	44.96	56.47	87.54	79.93
			5.79	6.81	7.71	6.57
10	85	0.2886+8	35.82	122.11	122.13	92.87
			10.8	9.27	8.28	10.93
11	390	0.1566+8	31.17	92.93	99.09	76.79
			6.47	13.41	20.98	19.19
12	460	0.1196+8	27.5	49.24	107.87	103.74
			2.14	6.68	7.42	7.43
13	2268	0.1674+9	40.69	32.68	53.2	52.56
			18.94	19.44	34.61	32.57

Tabella 1. *Clustering* su tutti i canali

25 prove in totale. Le variazioni nei *clusters* finali ottenuti, non sono risultate apprezzabili, c'è quindi da ritenere la *clustering* della Tabella 1 abbastanza stabile e definita.

4. Riduzione della dimensionalità e valori remoti

La *clustering* dei dati di una scena Landsat può essere agevolata se si applicano ai dati stessi delle tecniche capaci di depurare l'insieme da tutto quello che è ridondante o che comunque può comportare effetti negativi sulla partizione dei dati. Da più parti (*e.g.* Bryant, 1979) si riconosce ad esempio che i livelli di radianza registrati nei canali 6 e 7 sono fortemente correlati. La formula (3.2) che sintetizza la logica di aggregazione ha il suo secondo termine il logaritmo del determinante della matrice di dispersione Σ_g dei canali nel *clu-*

ster g-esimo. Uno stretto rapporto di linearità implica un valore basso di $|\Sigma_g|$. Questo non si verifica allo stesso modo in ogni gruppo: in alcuni, la quasi singolarità di Σ_g, è più accentuata, particolarmente in quelli più numerosi. Tali gruppi tenderanno allora ad assorbire sempre più unità man mano che le iterazioni proseguono con la conseguenza finale che i *clusters* ad alta frequenza perderanno del tutto la loro identità e non saranno più confrontabili con le categorie esistenti sulla scena. E' necessario dunque fare in modo che la *clustering* dei dati avvenga dopo che i rapporti di linearità siano stati rimossi qualora risultino effettivamente presenti.

Un'altra difficoltà è data dalla presenza di *outliers* ovvero di livelli di radianza fuori dal comune rispetto agli altri registrati sulla scena. Una tale evenienza non è certo strana quando si trattano dati multispettrali tenuto conto del tipo di fonte che origina i dati, del loro particolare modo di rilevanza e dello stato decisamente tormentato delle aree campione esaminate nel corso della ricerca. I valori remoti inficiano la stima dei centroidi dei *clusters*. Nella (3.2) i centroidi sono definiti attraverso le medie aritmetiche di canale. La media aritmetica risente particolarmente dei valori più discosti e questo può spostare il centro dei clusters verso punti di bassa frequenza con conseguente aumento della variabilità nel cluster e con l'importo di tutti gli effetti deleteri che questo può implicare.

Sia il problema della dimensionalità effettiva dei dati Landsat che quello dei valori remoti sono stati affrontati individuando una trattazione comune: l'analisi delle componenti principali.

Riduzione della dimensionalità

L' idea base di questa analisi è di descrivere la dispersione di un insieme di N punti in uno spazio euclideo a p dimensioni introducendo un nuovo insieme di coordinate lineari ortogonali tali che la variabilità del dato insieme di punti rispetto a questo nuovo sistema sia in ordine di grandezza decrescente ad ogni nuovo asse che viene inserito. Algebricamente l'analisi delle componenti principali prevede il calcolo degli autovalori $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ e degli autovettori $\alpha_1, \alpha_2, \ldots, \alpha_p$ della matrice di dispersione Σ dei dati. Le componenti principali richieste sono date dalle combinazioni lineari $\alpha_1^t X, \ \alpha_2^t X, \cdots \alpha_p^t X$ dove $\boldsymbol{X}$ in questo caso va intesa come la matrice *(p x N)* delle osservazioni ed α_i è l'autovettore normalizzato associato all'autovalore i-esimo di Σ. La riduzione della dimensionalità avviene, operativamente, scegliendo per la definizione del sottospazio in cui si proiettano le osservazioni originarie solo le prime $q < p$ componenti principali. L'indice q può essere determinato valutando la quantità

$$\frac{\sum_{i=1}^{q} \lambda_i}{Tr(\boldsymbol{\Sigma})} * 100 \qquad (4.1)$$

Quando la percentuale di variabilità spiegata dalle prime *q* componenti principali supera il 90% allora *q* è dato dalla (4.1).

Nella tabella 2, in due parti, vengono riportate le statistiche afferenti ai primi due momenti dei canali spettrali: da notare la similarità delle medie, la deviazione standard bassa nel canale 4 e quella relativamente alta nel 6.

	Canale 4	Canale 5	Canale 6	Canale 7
Medie	35.0112	37.5297	47.3253	0.6742
Dev.Std	24.6259	31.5367	46.8235	44.3451
Coeff.Var.	0.7034	0.8403	0.9894	0.9912

Tabella 2a. Medie deviazioni standard nella scena

	Canale 4	Canale 5	Canale 6	Canale 7
Canale 4	1	0.46463	0.66244	0.67422
Canale 5		1	0.80729	0.67367
Canale 6			1	0.95459
Canale 7				1

Tabella 2b. Matrice di correlazione sulla scena

La matrice di correlazione conferma l'esistenza di rapporti lineari accentuati fra canali, particolarmente fra quelli 6 e 7 mentre il canale 4 appare del tutto incorrelato con gli altri canali.

Nella Tabella 3 sono esposti i risultati dell'autoanalisi sulla matrice di correlazione.

Determinate della matrice di correlazione:0.011937					
Autovalori	3.14130	0.55036	0.28483	0.02354	
Var.spiegata	78.53%	92.29%	99.41%	100.00%	
	Varianze		Autovettori		
Channel 4	606.43995	0.439702	-0.783309	0.439389	0.005237
Channel 5	999.56345	0.470180	0.603871	0.603348	0.224149
Channel 6	2192.44020	0.550337	0.143410	-0.285875	-0.771259
Channel 7	1966.48790	-0.531717	-0.034660	-0.600986	0.595728

Tabella 3: esito dell'analisi dell componenti principali

La variabilità misurata lungo la prima componente principale è il 78,53% di quella totale e la variazione misurata lungo la seconda componente principa-

le è il 23,76%. Insieme le prime due componenti assorbono il 92,29% della varianza totale. Sembra quindi plausibile valutare a due la dimensione effettiva dei dati Lansat; Vale a dire che

$$\begin{aligned} Pc_1 &= 0.440C_4 + 0.470C_5 + 0.550C_6 - 0.532C_7 \\ Pc_2 &= -0.783C_4 + 0.604C_5 + 0.143C_6 - 0.035C_7 \end{aligned} \tag{4.2}$$

sono sufficienti a descrivere i dati Landsat della scena in esame.

Uno dei problemi più stringenti nell'analisi delle componenti principali è che le combinazioni lineari Pc_1 e Pc_2 presentano difficoltà ad una loro interpretazione significativa nei terreni delle variabili originarie.
Nella Tabella 4 è riportata la matrice di correlazione delle componenti principali e i valori dei 4 canali.

	Pc1	Pc2	Pc3	Pc4
Channel 4	-0.0794	-0.1405	-0.0561	0.0999
Channel 5	0.9427	0.9890	0.9680	-0.8882
Channel 6	0.9959	0.9632	0.9698	-0.9948
Channel 7	0.9868	0.9352	0.9679	0.9958

Tabella 4: correlazioni tra canali e componenti

Questo è di solito un passo che aiuta a dare un senso più preciso alla 4.2 In questo caso si riesce a vedere solo che il canale 4 ha dato scarso contributo alla definizione di Pc_1 e Pc_2 che invece sembra risentire in maniera abbastanza uniforme degli altri tre canali.

Nella figura 1 viene data una rappresentazione grafica degli autovettori. I "pesi" dei quattro canali nella prima componente suggeriscono in sostanza che la Pc_1 rappresenta un quadro generale delle intensità della radianza su tutte le bande. La seconda componente presenta invece dei pesi alternati in valore e grandezza, in altre parole la Pc_2 è la sintesi di una fattorizzazione bipolare dei canali. Lo studio in (Donker, 1976) trova risultati del tutto simili a quelli qui ottenuti e porta all'interpretazione della Pc_1 come la componente "*black and white*" e la Pc_2 come la componente "*colour*". Ulteriori chiarimenti in questo senso si possono vedere i lavori citati in (6). Dal punto di vista dell'analisi quantitativa dei dati Landsat quello che più conta è la possibilità di comprimere i dati originariamente in R4 con vantaggi sia pratici in quanto i tempi di esecuzione e lo spazio di memoria di classificazione ora occupano, sia teorici in quanto vengono di molto attenuati i problemi comportati dalla quasi singolarità delle matrici di dispersione di *clusters*.

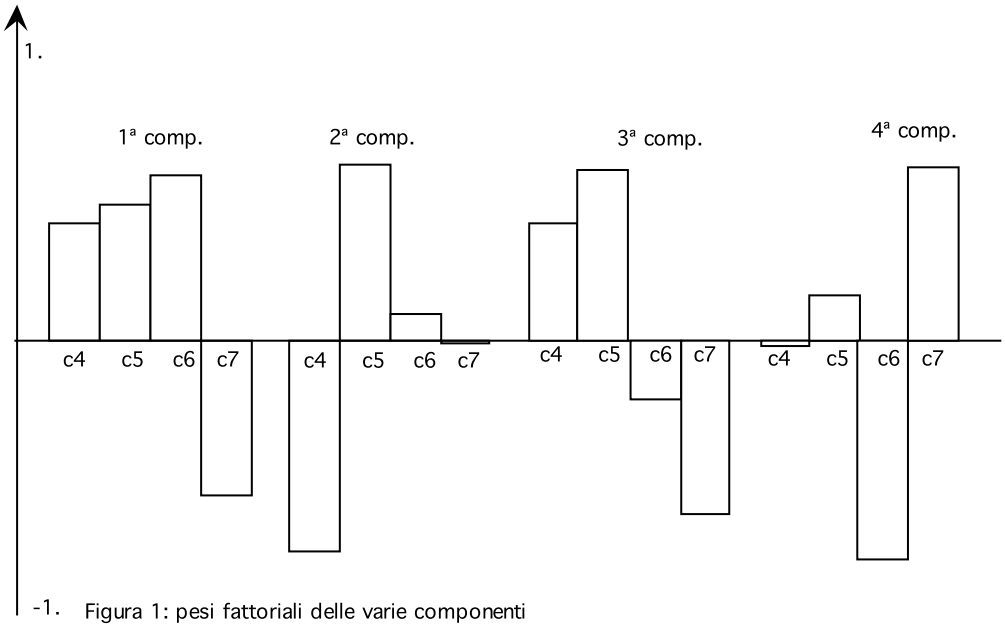

Figura 1: pesi fattoriali delle varie componenti

Nella Tabella 5 viene esposto il risultato finale dello stesso schema di *clustering* del terzo paragrafo applicato però alle Pc_1 e Pc_2.

Cluster	N. di Pixels	I ∑ I	Pc1		Pc2	
1	690	0.8377E + 6	29.11	3.41	39.06	2.99
2	471	0.10583 + 6	41.14	7.30	21.23	2.73
3	833	0.9065E + 7	45.57	5.38	30.17	6.09
4	1023	0.7856E + 7	51.75	6.88	30.17	8.03
5	408	0.7198E + 6	47.50	1.15	43.53	4.97
6	366	0.6881E + 5	27.81	1.58	34.65	5.97
7	523	0.7804E + 5	28.14	1.58	20.73	1.68
8	495	0.8228E + 6	28.51	1.64	41.63	5.06
9	402	0.8622E + 6	32.90	4.89	44.90	6.42
10	83	0.1048E + 7	35.81	4.36	32.36	6.58
11	158	0.3083E +5	29.98	3.94	35.60	6.31
12	607	0.4048E + 6	17.21	1.87	41.98	6.35
13	1632	0.8929E + 7	23.52	9.61	13.82	7.59
14	329	0.6977E + 5	42.71	5.27	45.29	10.06

Tabella 5: clustering in base alle prime deu componenti

I valori numerici che si riscontrano per la definizione dei 14 *clusters* sono ora più difficili da interpretare rispetto a quelli della Tabella 1. In questo senso è allo studio la predisposizione di una traccia interpretativa che parta dall'individuazione di un moderato numero di pixel della scena dei quali sia noto l'ente al terreno ad essi associato per poi estendere a tutta la scena i valori in Pc_1 e Pc_2. registrati in questi punti noti.

Valori remoti

Se per la riduzione delle dimensionalità si è portati a dare maggiore rilevanza alle componenti derivate dagli autovalori più grandi di $\sum$, nella individuazione degli *outliers* l'attenzione si concentra soprattutto sugli autovalori più piccoli. Le prime *q* componenti principali risentono infatti in maniera più acuta degli *outliers* che comportano di solito valori inappropriatamente grandi negli elementi di $\sum$. Gli autovalori $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ possono essere anche intesi come le varianze delle componenti e l'esame delle proiezioni delle variabili originarie lungo questi assi dove la variabilità è minore dovrebbe rendere più semplice il compito di studiare le deviazioni più palesi che si verificano.

Il metodo che si è adottato per lo studio degli *outliers* ricalca quello proposto in (Rao, 1964) e che impone la considerazione della somma degli scarti quadrati delle variabili originarie sulle ultime (p-q) componenti principali. In particolare si considera

$$\delta_i = \sum_{j=p-q+1}^{p} \left[a_j^t \left(X_{i-\mu} \right) \right]^2, \ i = 1, 2, \ldots, N \tag{4.3}$$

dove μ è il vettore delle medie su tutte le N osservazioni. Il metodo si articola in passi che prevedono prima la determinazione del massimo dei δ_i poi la definizione delle quantità

$$d_i = \frac{\delta_i}{Max\left\{ \delta_i | i = 1, 2, \ldots, N \right\}}, \ i = 1, 2, \ldots, N \tag{4.4}$$

Si ripartisce poi l'intervento unitario in un opportuno numero di subintervalli e si assegna ogni d_i all'intervallo entro cui il range è compreso. L'esame della distribuzione di frequenza dei d_i sui subintervalli guardando con particolare attenzione agli intervalli che comprendono un numero piccolo di punti dovrebbe alla fine consentire la localizzazione dei valori remoti.

Più che un metodo è una traccia intuitiva che può dare spazio ad errori ed approssimazioni, ma tenuto conto della difficoltà di studiare gli *outliers* nello spazio a più dimensioni (si pensi a come definire in maniera "oggettiva" un *outlier*) e della mole di dati da trattare vale la pena di prendere in considerazione. Nella tabella 6 viene riportata la distribuzione di frequenza relativa alla partizione dell'intervallo unitario in 20 subintervalli. Un fatto abbastanza evidente è che gli *outliers* possono essere localizzati in quei pixel che afferiscono alle classi con $d_i \geq 0.15$. Si tratta di 59 pixels (circa il 7%) che in questo caso, per la ridotta numerosità non possono influenzare la classifi-

cazione (il range dei valori di canale è in ogni caso limitato all'intervallo (0,255)) Indicando comunque la validità di questo tipo di tecnica quando si prendano in considerazione scene più estese.

Intervalli	N. di pixels	Intervalli	N. di pixels
0,00 - 0,05	4743	0,50 - 0,55	0
0,05 - 0,10	4202	0,55 - 0,60	3
0,10 - 0,15	1005	0,60 - 0,65	1
0,15 - 0,20	0	0,65 - 0,70	4
0,20 - 0,25	0	0,70 - 0,75	0
0,25 - 0,30	14	0,75 - 0,80	1
0,30 - 0,35	8	0,80 - 0,85	0
0,35 - 0,40	9	0,85 - 0,90	0
0,40 - 0,45	11	0,90 - 0,95	1
0,45 - 0,50	3	0,95 - 1,00	1

Tabella 6: distribuzione dei valori remoti

5. Conclusioni

In questo lavoro si sono mostrate alcune tecniche di analisi multivariata applicate allo studio dei dati MSS e di come queste possono essere utilizzate per agevolare l'interpretazione visuale di immagini Landsat. In particolare si è tentata una sistematizzazione teorica più precisa della tecnica di *clustering* I.S.O.D.A.T.A., d'elezione nell'approccio *unsupervised* al *remote sensing*, ripensandola nei termini dell'analisi di *mixtures* di distribuzioni normali multivariate. Si è anche dato risalto a due dei problemi del *remote sensing* alla cui soluzione i modelli statistici danno un contributo cospicuo: la dimensionalità effettiva dei dati Landsat e la presenza di valori remoti.

Le semplificazioni che risultano dall'applicazione dei modelli statistici possono sicuramente agevolare chi poi deve, di fatto, valutare le immagini ricostruite in base ai livelli di radianza. Non possono però dare la chiave di soluzione di tutti i problemi del *remote sensing* che resta invece vincolata alla interdisciplinarietà ed alla iteratività di cui si è parlato nell'introduzione.

Bibliografia

Ball G.H. Hall D.J. (1967). A clustering technique for summarizing multivariate data. *Behavioral Science,* 12, 153-155.

Bryant J. (1979). On the clustering of multidimensional pictorial data. *Pattern Recognition,* 2, 115-125.

Cormack R.M.(1971). A review of classification. *Journal of the Royal Statistical Society, A,* 134, 32-55.

Curtis L.F. (1978). Remote sensing systems for monitoring crops and vegetations. *Progress in Phisical Geography,* 2, 55-79.

Day N.E. (1969). Estimating the components of a mixtures of normal distributions. *Biometrika,* 56, 463-474.

Donker N.H.W. Mulder N.J.(1976). Analysis of MSS digital imagery with the aid of principal transforms. Proceedings of the XIII congress of the international society for photogrammetry. Helsinky.

Follestad B.A. Levandosky D.W. (1976). Analysis of Landsat data for mapping superficial deposits.Proceedings of the XIII congress of the international society for photogrammetry. Helsinky.

Gower J.C. (1967). A comparison of some methods of cluster analysis. *Biometrics,* 23, 623-634.

Huddlestone H.F. Wighton W.H.(1975). Use of remote sensing in sampling agricultural data. *Bulletin of the International Statistical Institute,* 105-119.

Mejer J. Baldi G. Cellerino G. De Carolis C. La Pietra G. (1976). Classification automatique de données du satellite Landsat appliquée à agricolture et sylviculture. *Il Riso,* 24, 291-303.

Rao C.R. (1964). The use and interpretarion of principal components in applied research. *Sankhya, A,* 26, 349-358

Scott A.J. Simons M.J. (1971). Clustering methods based on likelihood ration criteria. *Biometrics,* 27, 387-397.

Wolfe J.H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research,* 5, 329-341.